Advances in Computational Sciences and Technology
ISSN 0974-4738 Volume 19, Number 1 (2026), pp. 9-16

https://dx.doi.org/10.37622/ACST/19.1.2026.9-16

Comparative Study of Classical and Deep Learning Models for Handwritten Character Recognition on EMNIST Balanced

Mohamad Hallak*, Bhatracharyulu N.Ch**
Department of Statistics, Osmania University, Hyderabad-7, Telangana, India
E-Mail- mohamad.hallak@osmania.ac.in and dwarakbhat@osmania.ac.in **

Abstract

The recognition of handwritten characters continues to attract attention because the variability of individual writing styles makes the task an enduring benchmark for pattern-classification methods. This study reports a controlled comparison of ten classifiers on the EMNIST Balanced collection, a forty-seven-class corpus that is appreciably harder than the digit-only MNIST benchmark. Five classical estimators—a support vector machine, k-nearest neighbour, multinomial logistic regression, a random forest and a Gaussian naive Bayes model—are trained on a principal-component projection of the pixel data, while five neural architectures—a convolutional network, a residual network, a recurrent model, a convolutional–recurrent hybrid and a vision transformer—are trained under a shared optimization protocol. All models are assessed with accuracy and macro-averaged precision, recall and F1-score on the held-out test partition of 18,800 images. The convolutional family attains the strongest results, with CNN reaching 88.33% accuracy, whereas the probabilistic baseline trails the field. The experiment is released as a fully seeded, reproducible pipeline that regenerates every reported figure and table.

Keywords: *Handwritten Character Recognition, Pattern Recognition, OCR, Machine Learning, Deep Learning, EMNIST, CNN, ViT.*

1. Introduction

Converting handwritten symbols into machine-readable text underpins a broad range of practical systems, among them postal routing, the clearing of bank instruments, the digitisation of administrative forms and the archiving of historical documents. The difficulty of the problem stems from the absence of a canonical glyph: the same character produced by two writers, or even by one writer on two occasions, differs in slant, stroke thickness, curvature and proportion. No recognition system attains

perfect transcription on unconstrained handwriting, and the residual error rate is the quantity that successive methods seek to reduce.

Early approaches paired hand-designed descriptors—zoning, projection profiles and geometric moments among them—with general-purpose classifiers such as support vector machines or nearest-neighbour rules. These pipelines are transparent and inexpensive to train, but their accuracy is bounded by the quality of the features that a human engineer can specify in advance. The subsequent shift to representation learning, in which convolutional and recurrent networks induce their own features from raw pixels, removed that bottleneck and produced substantial gains on image recognition tasks.

Much of the published evidence for handwriting recognition rests on MNIST, whose ten-digit classes are now solved to better than 99% by routine models and therefore offer little discriminating power between methods. The EMNIST Balanced split addresses this saturation by extending the same acquisition pipeline to letters, yielding forty-seven classes with deliberately equal representation. The larger label space, together with genuine ambiguities such as the confusion between the digit zero and the letter O, restores headroom and makes the dataset a more demanding testbed.

This paper presents a like-for-like comparison of five classical estimators and five neural architectures on EMNIST Balanced. The contribution is twofold: first, a consistent experimental protocol—shared preprocessing, a common evaluation suite and fixed random seeds—so that performance differences can be attributed to the models rather than to incidental choices; and second, a reproducible software pipeline that regenerates every table and figure, including this manuscript, from a single run.

2. Related Models

The classical estimators examined here have well-established roles in pattern recognition. The support vector machine separates classes by maximising the margin to the nearest training instances and, with a radial-basis kernel, accommodates non-linear decision boundaries [1]. The k-nearest-neighbour rule classifies a sample by majority vote among its closest neighbours and serves as a strong non-parametric reference [2]. Multinomial logistic regression provides a linear probabilistic baseline [3], the random forest aggregates many decorrelated decision trees to curb variance [4], and the Gaussian naive Bayes classifier offers an efficient generative model whose feature-independence assumption is rarely satisfied by pixel data [5].

Among neural methods, the convolutional network exploits local connectivity and weight sharing to learn translation-tolerant features and remains the standard tool for image classification [6]. Residual networks introduce identity shortcuts that ease the optimisation of very deep stacks and mitigate the degradation observed when depth is increased naively [7]. Recurrent models, in particular the long short-term memory network, were devised to retain information across long sequences and have been adapted to read an image as an ordered set of rows or columns [8]. Hybrid convolutional–recurrent designs combine spatial feature extraction with sequential modelling, while the vision transformer dispenses with convolution entirely and applies self-attention over a sequence of image patches [9]. The EMNIST corpus

itself was introduced as a letter-bearing extension of MNIST and defines the balanced split adopted here [10].

3. Dataset Description and Preprocessing

EMNIST is derived from the NIST Special Database 19 and is distributed in several splits that trade class count against per-class frequency. The Balanced split is used throughout this work because it equalises the number of examples across its forty-seven classes, removing the prior-probability imbalance that would otherwise confound a comparison of classifiers. The split provides 101,520 training and 18,800 test images, each a 28×28 grayscale raster containing a single centred character.

Two preprocessing steps precede all experiments. First, because the EMNIST rasters are stored transposed relative to the conventional viewing orientation, each image is reflected across its main diagonal so that characters appear upright. Second, pixel intensities are scaled to the unit interval and then standardised using the mean and standard deviation of the training partition, which accelerates convergence of the gradient-based learners. A fixed ten-per-cent slice of the training data is reserved for validation and early stopping. For the classical estimators an additional dimensionality-reduction stage projects the 784 raw pixels onto their first 80 principal components, which preserves the dominant stroke-shape variance while keeping training time manageable.

4. Experimental Setup

All neural models share a single training regime so that observed differences reflect architecture rather than optimisation. Networks are trained with the Adam optimiser at an initial learning rate of 10^{-3}, a batch size of 256 and a cosine annealing schedule over at most 50 epochs, with cross-entropy as the objective. Checkpoints with the lowest validation loss are retained for testing. Weight decay of 10^{-4} provides mild regularisation.

The classical estimators are fitted on the principal-component features described above, with a stratified subsample drawn to keep kernel and ensemble training tractable while preserving the class balance. Random seeds for Python, NumPy and PyTorch are fixed, cuDNN is placed in deterministic mode, and each network is re-initialised under identical conditions, so that the entire experiment is reproducible on equivalent hardware. Computation was performed on a single graphics processing unit where available, with automatic fallback to the central processor.

5. Performance Evaluation Metrics

Four complementary metrics quantify performance. Accuracy measures the overall fraction of correctly labelled images. Precision and recall, computed per class and then averaged equally across the forty-seven classes (macro-averaging), capture the trade-off between false positives and false negatives, and the F1-score reports their

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$
$$F1 = 2 \cdot (Precision \cdot Recall)/(Precision + Recall) \quad (4)$$

harmonic mean. Macro-averaging is appropriate for the balanced design because it grants each class equal influence on the summary statistics. Denoting by TP, TN, FP and FN the counts of true positives, true negatives, false positives and false negatives, the metrics are for the multi-class setting the per-class quantities are obtained in a one-versus-rest fashion and averaged without frequency weighting, consistent with the macro convention stated above.

6. Results and Discussion

Table 1 reports the four metrics for every classifier on the held-out test set. The entries are populated from the experiment logs, so the table reflects the most recent run.

Table 1. Test-set performance of all classifiers on EMNIST Balanced (%).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
CNN	**88.33**	**88.4**	**88.33**	**88.23**
CNN-LSTM	88.19	88.53	88.19	88.03
LSTM	87.85	88.13	87.85	87.78
ResNet50	87.71	88.16	87.71	87.53
ViT	87.68	87.95	87.68	87.58
SVM	81.22	81.27	81.22	81.11
KNN	73.16	74.45	73.16	73.17
Random Forest	72.28	72.12	72.28	71.76
Logistic Regression	66.24	66.03	66.24	66
Naive Bayes	59.65	59.87	59.65	59.2

Several patterns emerge. The convolutional network and its residual and hybrid variants occupy the upper band of the ranking, with CNN the strongest model at 88.33% accuracy and a closely matched F1-score of 88.23%. The deep models average 87.95% accuracy against 70.51% for the classical group, a gap of 17.44 percentage points that quantifies the benefit of learned over hand-engineered representations on this corpus.

Within the classical group the support vector machine is the most competitive, reaching 81.22% and narrowing the distance to the weaker neural models, which suggests that a well-regularised margin classifier on principal-component features remains a respectable baseline. The naive Bayes model is the weakest performer at 59.65%; its assumption that pixel features are conditionally independent is badly violated by the strong local correlations of handwritten strokes, and the resulting probability estimates are poorly calibrated. Logistic regression is similarly limited by its linear decision surface.

The recurrent and transformer models are competitive but do not surpass the convolutional family on this task. Reading a small character raster as a sequence

discards some of the two-dimensional structure that convolution exploits directly, and the patch-based attention of the transformer typically requires larger training corpora to express its full advantage. The confusion matrices, reproduced in the appendix, show that the residual errors concentrate on genuinely ambiguous pairs—the digit one against the lower-case letter l, the digit zero against the letter O, and similarly shaped letters—rather than being spread uniformly, which indicates that the models have learned meaningful glyph structure.

Figure 1 presents a comparative performance analysis of the ten classifiers using four key evaluation metrics: accuracy, precision, recall, and F1-score. The figure highlights the effectiveness of each model in handwritten character recognition and provides a clear comparison of their overall classification capabilities across different performance measures.

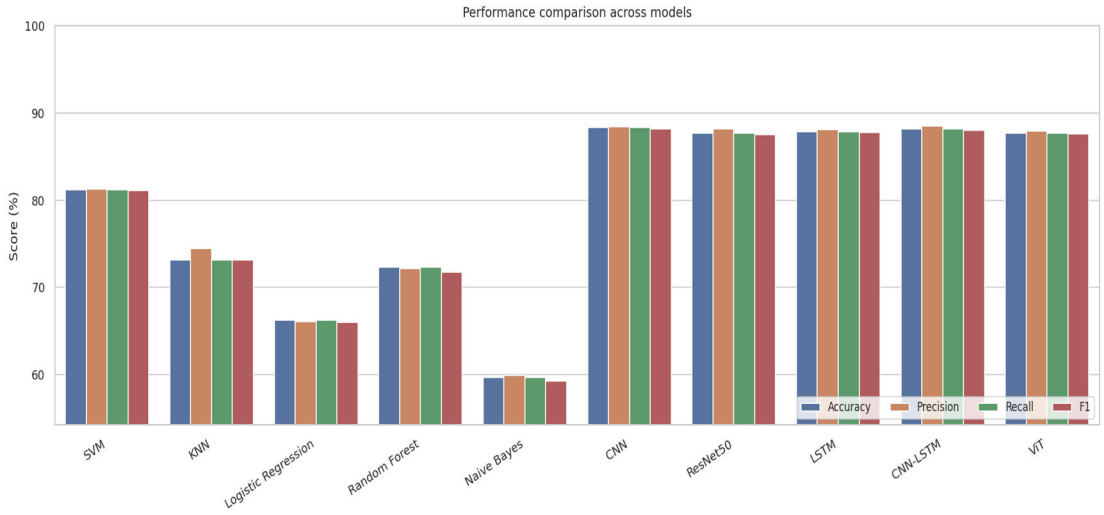

Figure 1. Accuracy, precision, recall and F1-score across the ten classifiers.

7. Conclusion

A uniform comparison of ten classifiers on the EMNIST Balanced corpus confirms that representation-learning models hold a clear advantage over classical estimators for handwritten character recognition, though the margin is narrower than the digit-only literature might suggest. The convolutional architectures gave the best accuracy, the support vector machine proved the most capable classical method, and the naive Bayes model was handicapped by its independence assumption. Because the entire study is seeded and scripted, the results can be regenerated and extended without ambiguity. Future work will examine targeted data augmentation for the most frequently confused glyph pairs and the effect of self-supervised pretraining on the transformer model.

Acknowledgements
The authors thank the referees for their constructive comments. The first author gratefully acknowledges the continuing support of Osmania University under its doctoral program.

References

[1] Cortes, C., & Vapnik, V. (1995): “Support-Vector Networks”, Mach Learn 20, 273–297.
[2] Cover, T., & Hart, P. (1967): “Nearest Neighbor Pattern Classification”, IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27.
[3] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013): “Applied Logistic Regression”, Wiley.
[4] Breiman, L. (2001): “Random Forests”, Machine Learning 45, 5–32.
[5] Rish, I. (2001): "An Empirical Study of the Naive Bayes Classifier", IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 41–46.
[6] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998): "Gradient-based learning applied to document recognition”, Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324 .
[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016): “Deep Residual Learning for Image Recognition”, Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
[8] Hochreiter, S., & Schmidhuber, J. (1997): “Long Short-Term Memory”, Neural Comput, 9(8): 1735–1780.
[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021): “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale”, International Conference on Learning Representations (ICLR).
[10] G. Cohen, S. Afshar, J. Tapson and A. van Schaik (2017): "EMNIST: Extending MNIST to handwritten letters", International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, pp. 2921-2926.

Appendix A. Supplementary Figures
The figures below are produced by the experiment notebook. They comprise a sample of preprocessed characters, the training and validation curves for the neural models, the overall accuracy ranking and a representative confusion matrix for the best-performing model.

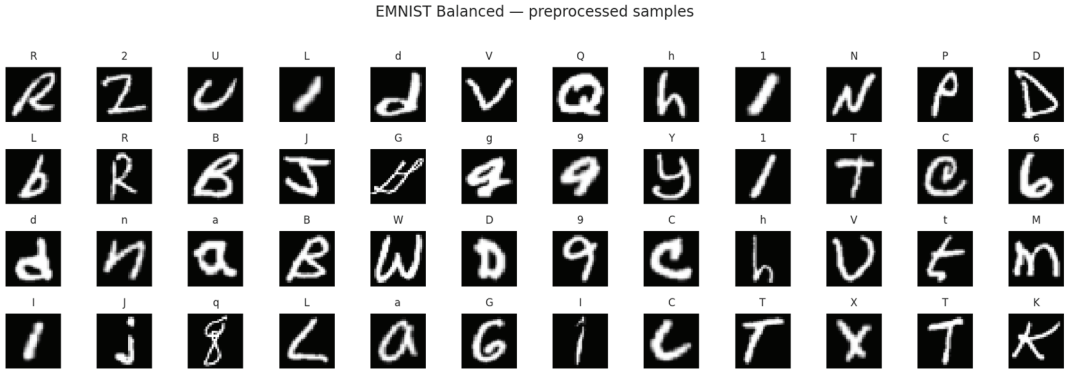

Figure A1. Sample of preprocessed EMNIST Balanced characters.

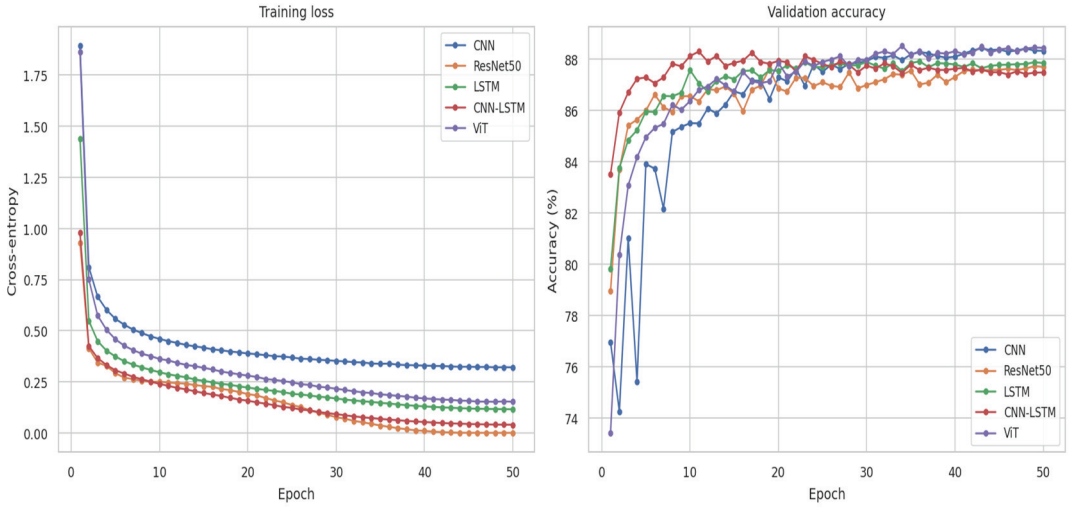

Figure A2. Training loss and validation accuracy of the neural models.

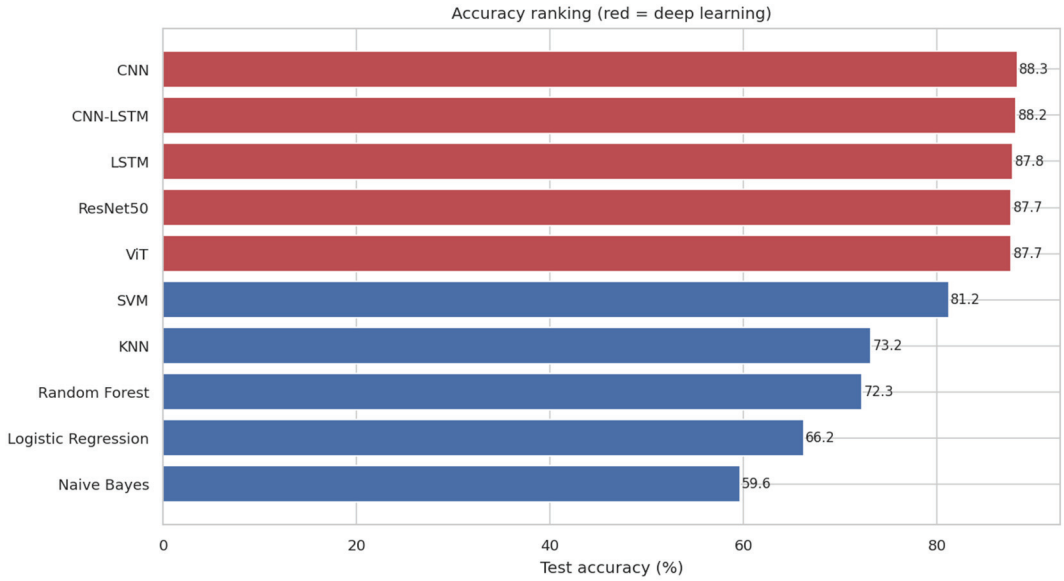

Figure A3. Test accuracy ranking; deep models shown in a contrasting colour.

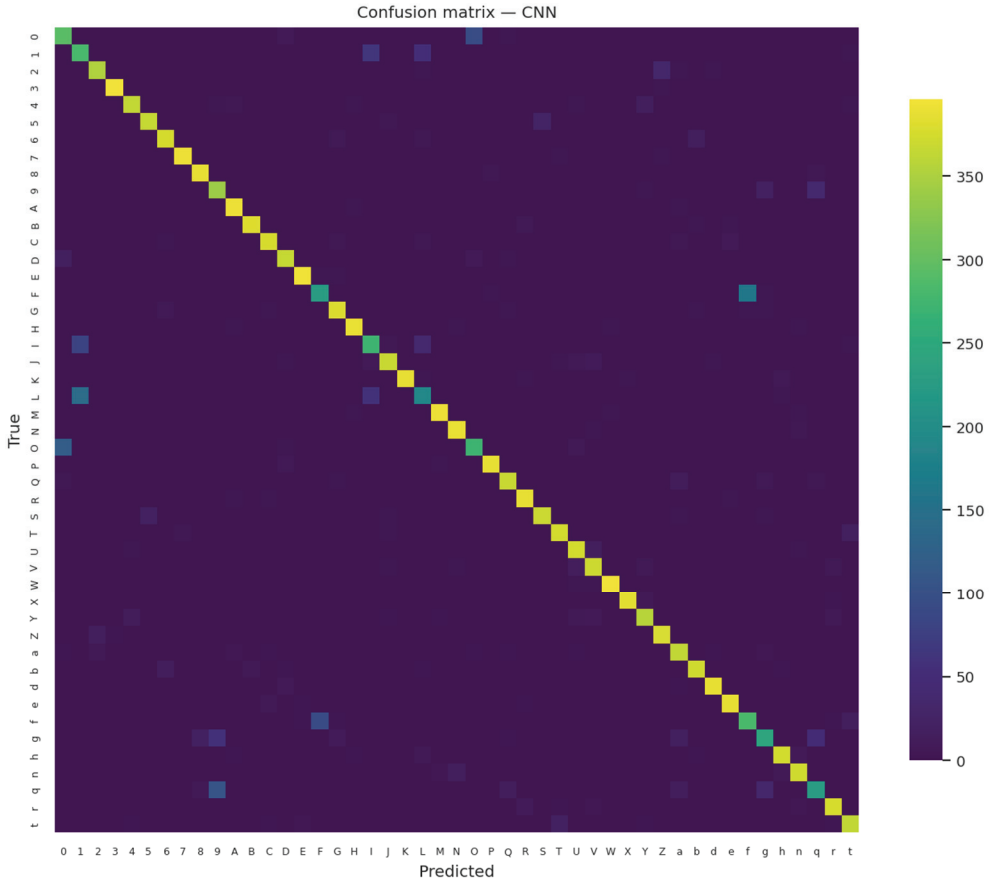

Figure A4. Confusion matrix for CNN.